Efficient metagenomic microbial community profiling using unique clade-specific marker genes

Nicola Segata[1], Levi Waldron[1], Annalisa Ballarini[2], Vagheesh Narasimhan[1], Olivier Jousson[2], Curtis Huttenhower[1]

[1] Department of Biostatistics, Harvard School of Public Health, Boston (MA), USA
[2] Centre for Integrative Biology, University of Trento, Trento, Italy

Supplementary Note 1. Limitations of current computational methods for microbial community profiling.

Current bioinformatic methods for metagenomic community profiling are limited by three main factors including computational expense, untenable accuracy for short (<400nt) reads, and difficulties in normalizing read-level relative abundances into per-organism relative abundances. Computationally, millions or billions of metagenomic reads classified by mapping must typically be compared to thousands of microbial reference genomes, making the performance of the algorithms employed for this alignment a bottleneck. In contrast, methods considering only sequence composition typically encounter neither of these efficiency issues, but frequently entail reduced accuracy and a loss of species-level resolution. Short read lengths, as obtained by Illumina sequencing, thus present particular difficulties for compositional methods, as precise and robust features become challenging to establish. Finally, all current methods provide clade assignments on a per-read basis, but they fail to normalize these assignments based on differences in organisms' genome sizes, gene copy number variations, and mapping confidence scores. Raw read-to-clade mappings are thus not a direct proxy for the cellular composition of a microbial community, regardless of the efficiency or accuracy of the computational method by which they are derived.

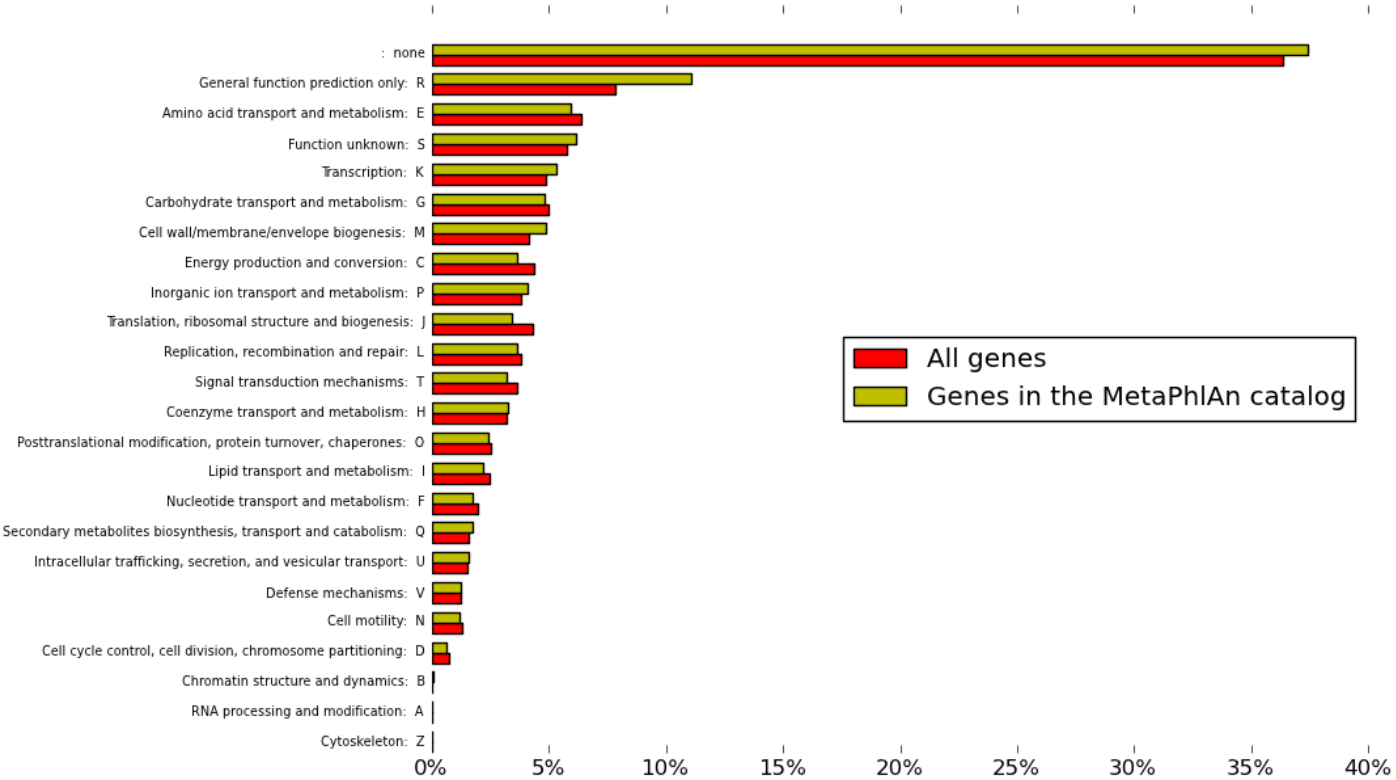
Supplementary Note 2. Enterotypes and species-level abundance patterns in the healthy gut microbiome.

MetaPhlAn's estimates of species-level abundance allowed us to refine the investigation of genus-level gut microbiome clusters, referred to as enterotypes[1] (Figure 3C). Enterotype 2 (*Prevotella*-dominant) remained clearly identifiable, but the *Bacteroides* of Enterotype 1 were instead diversified in a manner quite similar to lactobacilli in the vaginal microbiota, although with more species and less exclusive dominance. In particular, *B. ovatus, B. vulgatus*, and *B. stercoris* characterized three distinct sets of samples and, interestingly, *B. eggerthii* and a yet-to-be-sequenced *Bacteroides* species demonstrated discrete abundance patterns alternating dominance with near-complete absence. Other species achieving dominance in multiple samples included *Alistipes putredinis*, *Dialister invisus*, *Eubacterium siraeum*, *Eubacterium rectale*, and *Butyrivibrio crossotus*.

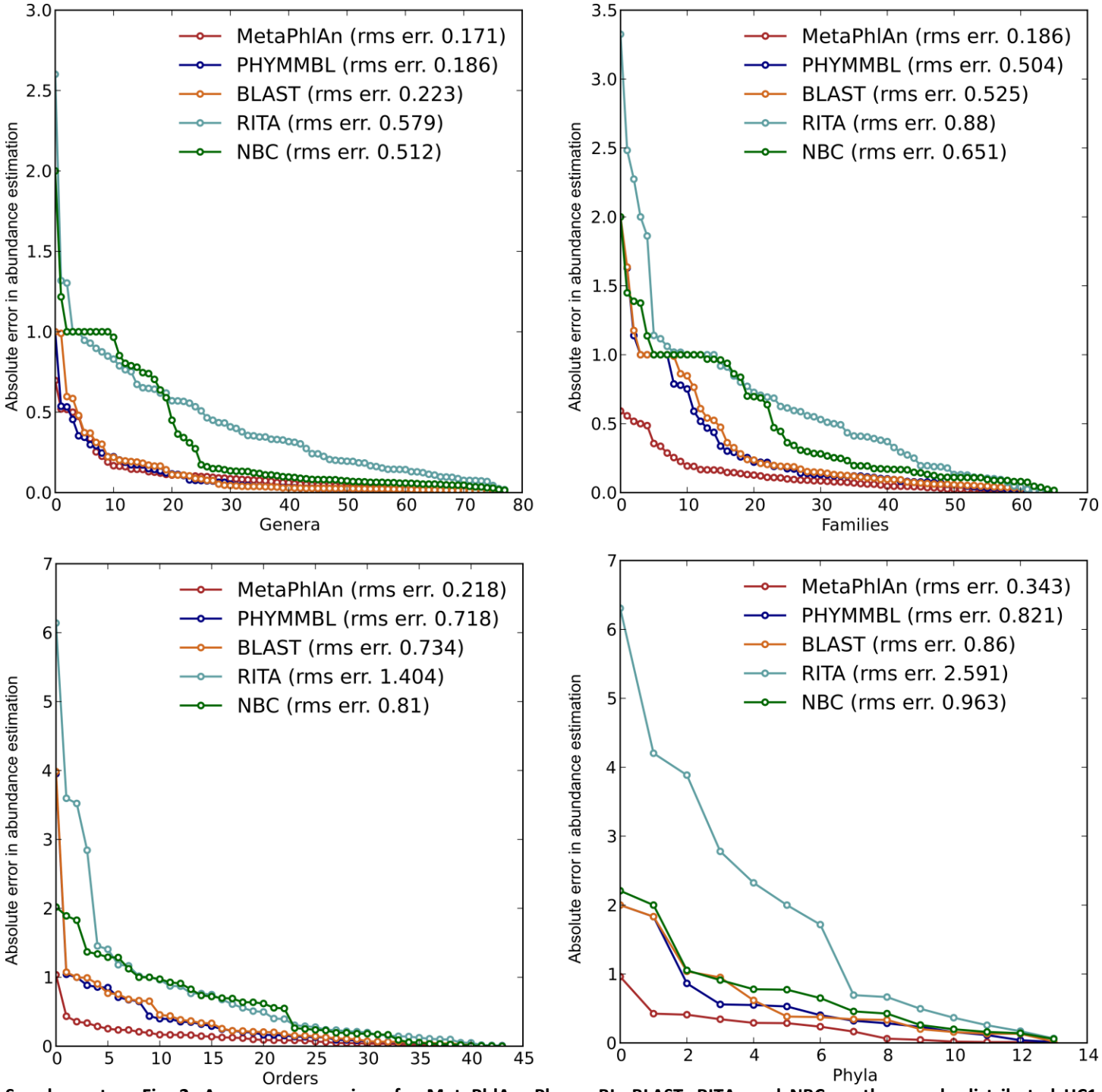
This species-level analysis together with the genus-level patterns observed in Fig. 3B and detailed in the main text, yielded a unique view of the presence of well-defined microbial community compositions or enterotypes. When investigating enterotypes, these data provide no clear support for a *Bacteroides* genus enterotype, as the abundance of the clade forms a continuous gradient throughout the samples without evidence of a discrete community type. Critically, however, at the species level, several small well-defined clusters were detected, suggesting a more tractably discrete organization of diversity below the genus level. This would argue that the healthy gut might behave more similarly to the healthy vaginal microbiota, in that detectable discrete community states should be sought at the species or strain level. In fact, the only genus-level cluster confirmed in these data was the *Prevotella* enterotype, which proved to consist of only one species (*P. copri*). Even at the species level, all of these data remain cross-sectional and not longitudinal; further investigations of the stability and reproducibility of microbiota types and their correlation with environmental factors will certainly be needed, and should be performed in terms of OTUs, species, or strains.

Supplementary Note 3. Combined analysis of microbiomes in distinct healthy populations.

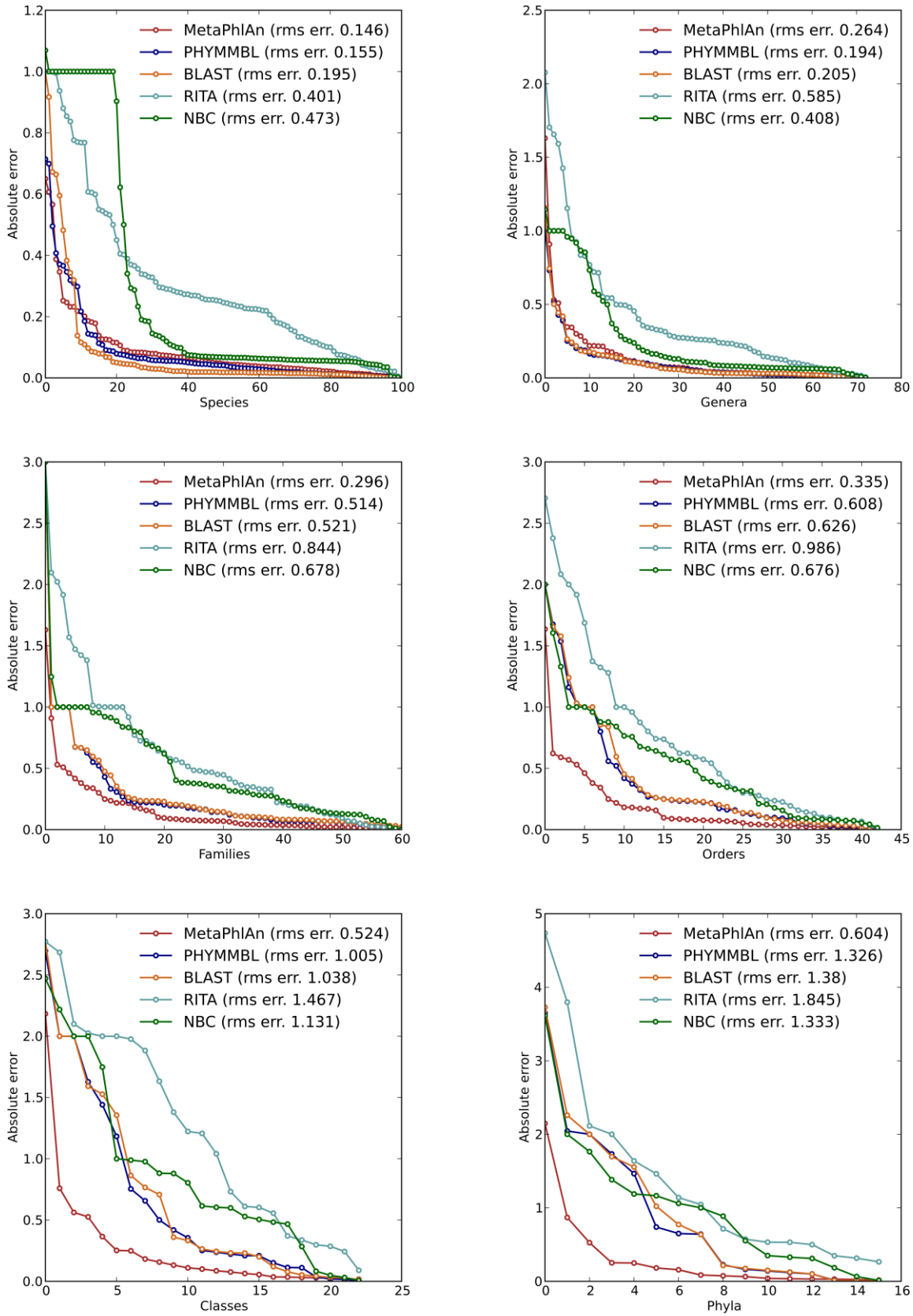
The clades most different between the HMP and MetaHIT samples were *Bacteroides ovatus* (5.8%±10.1% in HMP, 0.5%±0.9% in MetaHIT) and *Phascolarctobacterium* (0.13%±0.59% in HMP, 1.13±2.18 in MetaHIT). At higher taxonomic levels, samples from MetaHIT were on average enriched for Clostridia and Bifidobacteriales. Intriguingly, these differences were not best explained by shifts across all members of the two cohorts. Instead, they corresponded to differential representation of subjects with individually high or low populations of these bacterial species. Many HMP samples were dominated by specific *Bacteroides* species, for example, especially *B. ovatus* and *B. vulgatus*, whereas comparable numbers of samples dominated by *Prevotella* appeared in both cohorts. Other evenly represented clusters included those dominated by *Butyrivibrio crossotus* and by *Eubacterium rectale*. The fact that samples with highest proportions of all *Bacteroides* species were found in the American cohort is particularly interestingly, because this phenomenon has been recently associated with long-term diets rich in protein and animal fat[2]. In addition, MetaPhlAn highlighted the contribution of several different species to the overall abundance of *Bacteroides*, and likewise of a variety of species-level clades to the Firmicutes. This suggests that the correlation between microbiota and factors like diet or disease is driven by characteristics of the dominating species - or even strains - and not by the total impact of whole phylum or genus in the community composition.



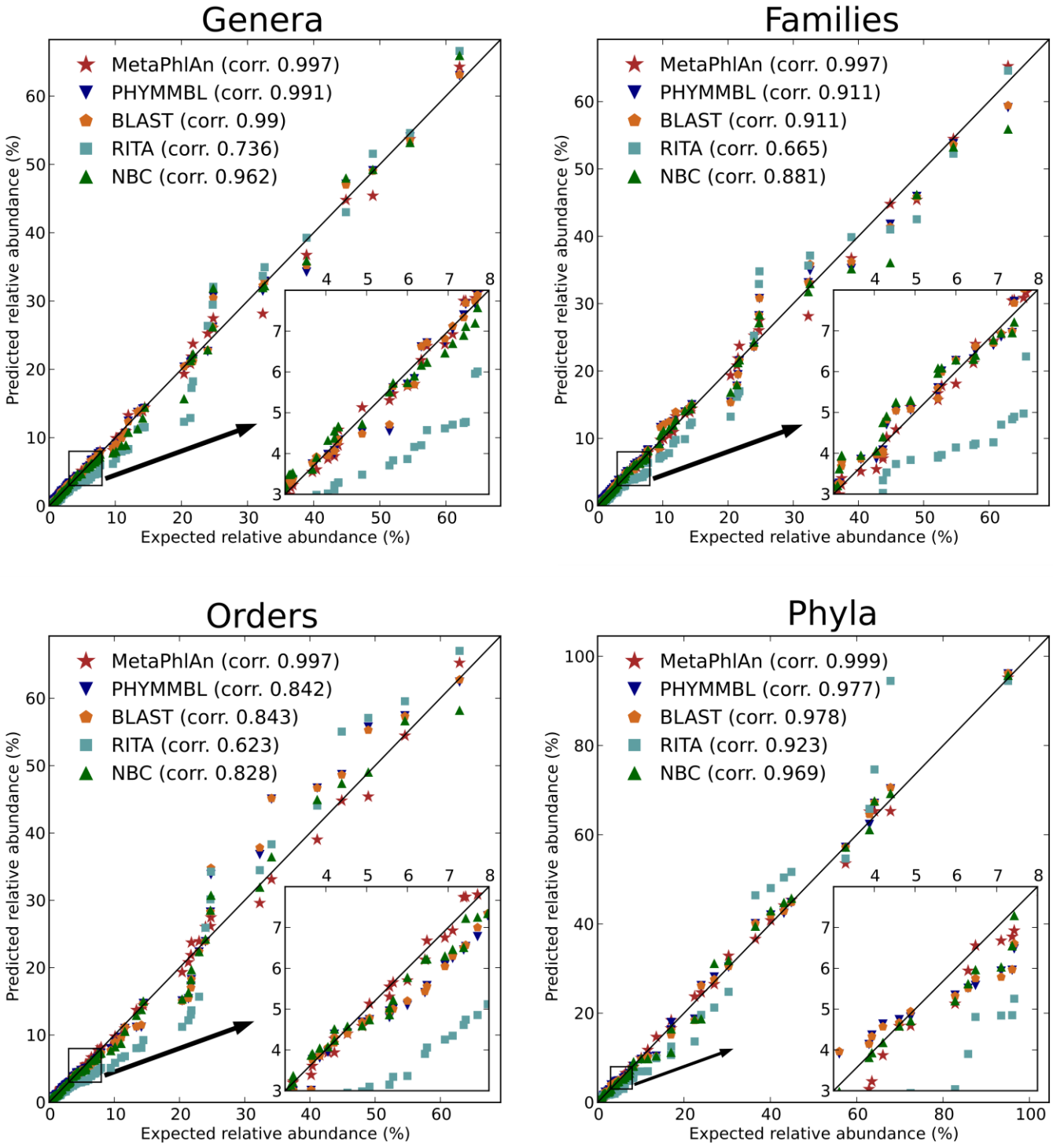
Supplementary Fig. 1: The MetaPhlAn marker database covers all functional modules with fractions comparable to the total background distribution computed using all available genomes. Yellow bars represent the fraction of MetaPhlAn marker genes in each high-level functional category in the COG database. The overall functional distribution of all the genes in IMG JGI is also reported as a reference background distribution.



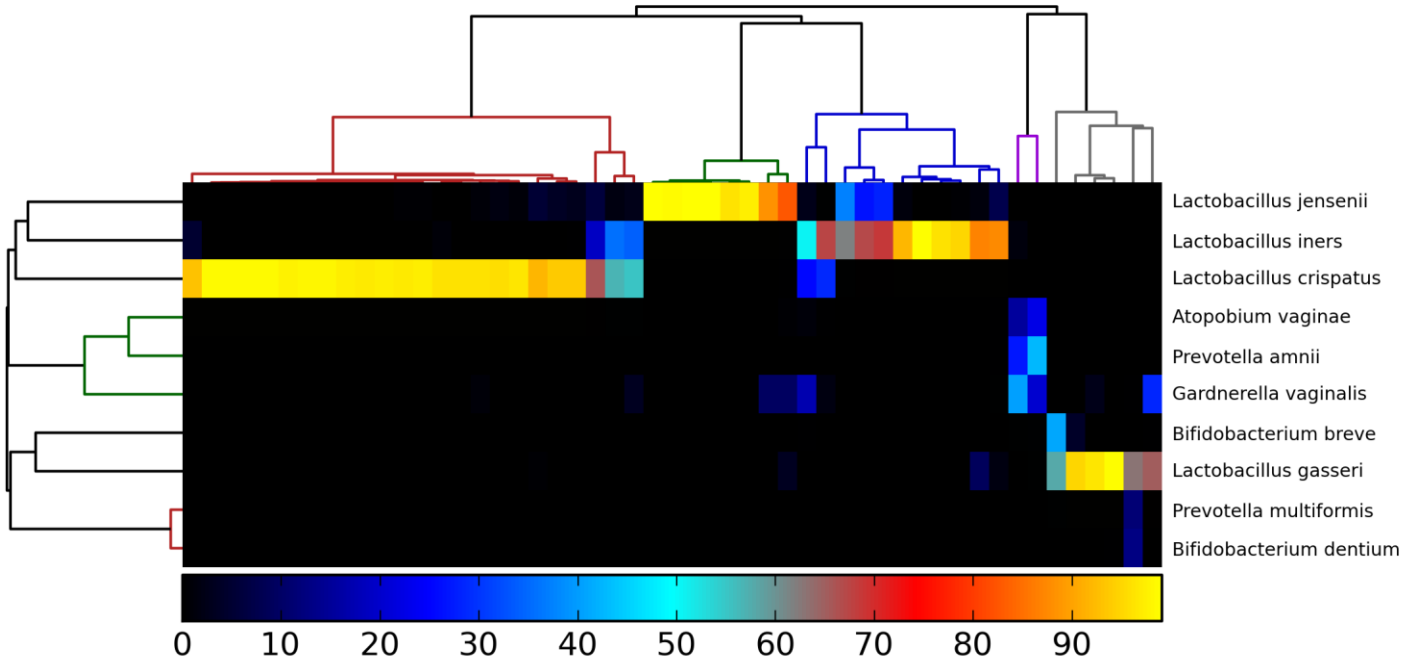
Supplementary Fig. 2. Accuracy comparison for MetaPhlAn, PhymmBL, BLAST, RITA, and NBC on the evenly distributed HC1 synthetic metagenome. The comparison detailed in Figure 1 A-B for species and classes is extended here to genera, families, orders and phyla. The HC1 metagenome is composed by 1,000,000 reads from 100 different organisms with identical relative abundances of genome copy numbers (1%).



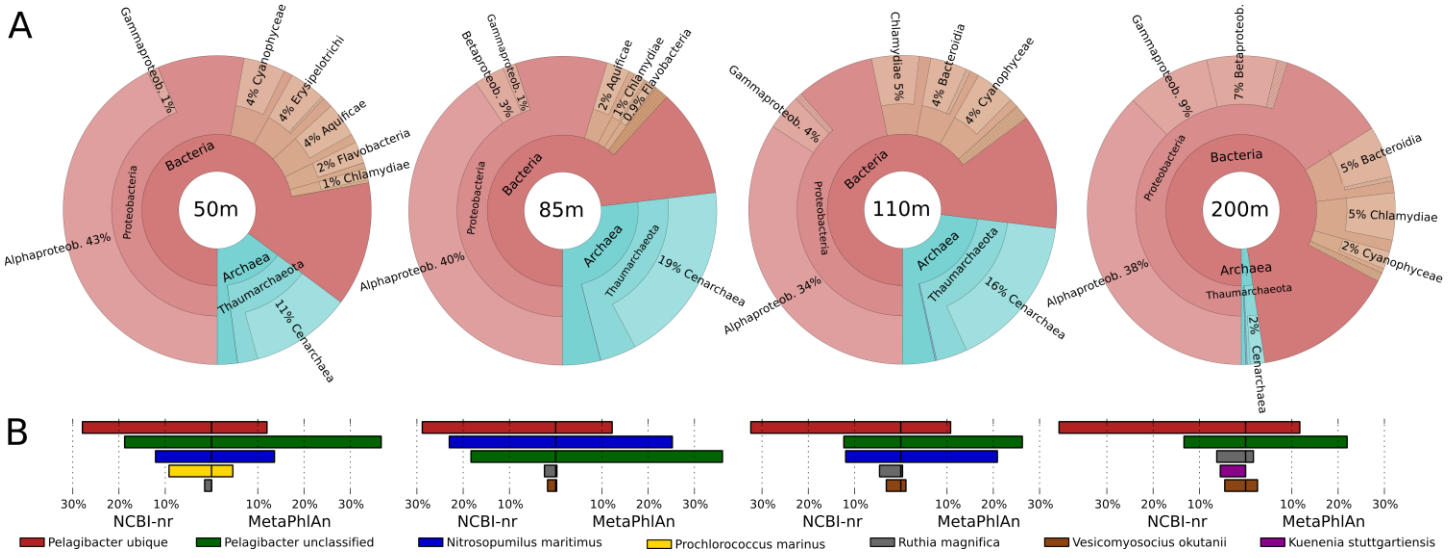
Supplementary Fig. 3. Accuracy comparison of several methods on the evenly distributed HC2 synthetic metagenome. MetaPhlAn, PhymmBL, BLAST, RITA, and NBC are compared on a second high-complexity metagenome (HC2) with evenly distributed abundances build using 100 organisms not considered in HC1. All taxonomic levels are reported here with the rooted mean squared errors of all the tested methods.



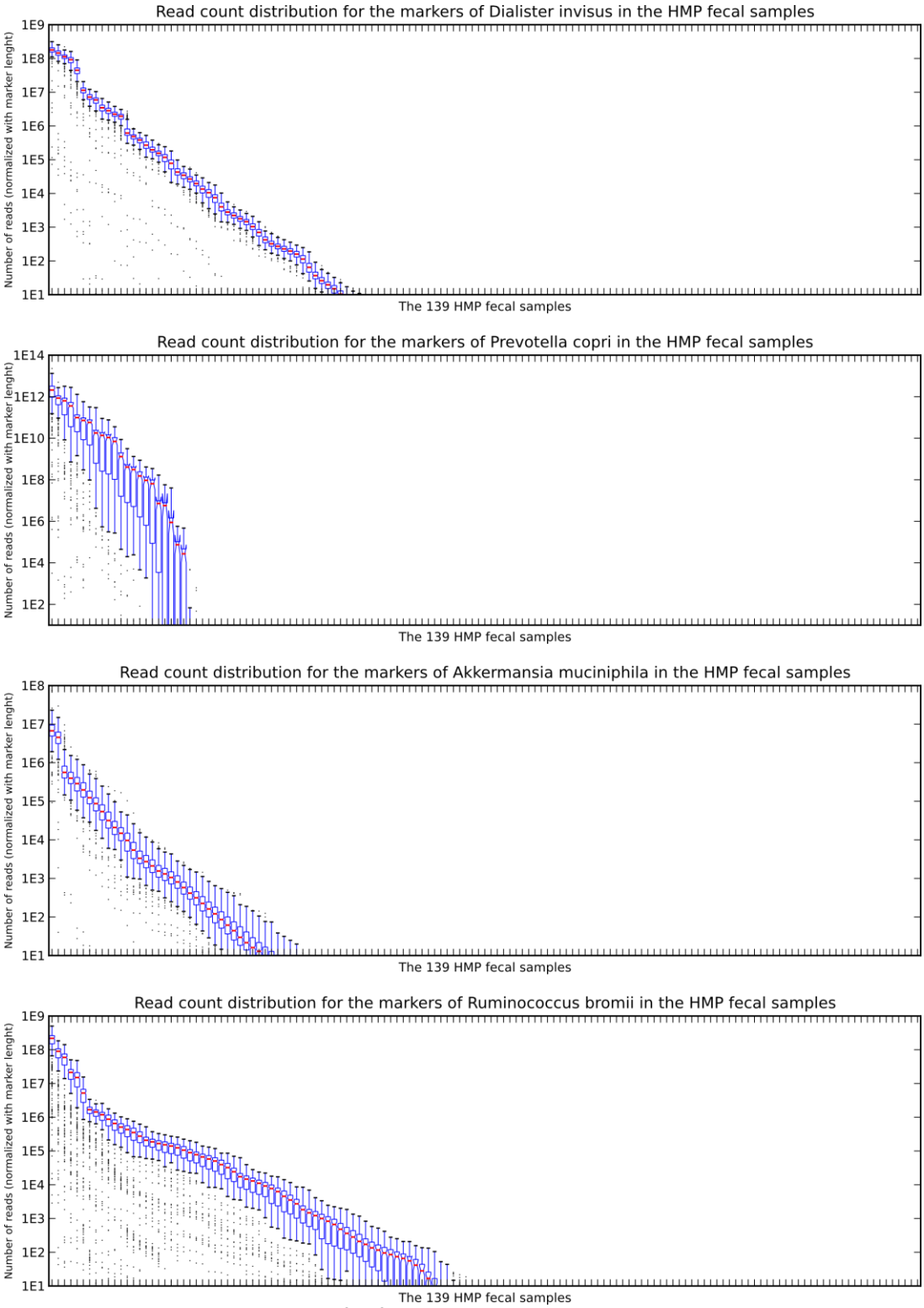
Supplementary Fig. 4. Comparison of MetaPhlAn with other computational methods in predicting the taxonomic composition of 8 synthetic metagenomes with staggered microbial abundances. We extend here the comparison described in Figure 1 C reporting the genus-, family-, order-, and phylum-level quantification of organismal abundance estimated from 8 low-complexity (25 organisms each) synthetic communities with log-normally distributed abundances.



Supplementary Fig. 5. Species-level hierarchical clustering for the HMP vaginal samples. The 51 posterior fornix samples from the HMP cohort were profiled by MetaPhlAn and the ten most abundant species hierarchically clustered (using Bray-Curtis similarity among samples and Pearson correlation among species). The five resulting clusters coincide with high abundances (above 50% percentage abundance) of four different *Lactobacillus* species and with their absence as reported previously on an independent dataset[3].



Supplementary Fig. 6. Community compositions on four marine metagenomes sampled at different depths in permanent marine oxygen minimum zones (data from Stewart et al.[4]). A) MetaPhlAn's community composition inferences from the domain to class level highlight a substantial presence of archaeal organisms (notably underrepresented in the deepest sample) and Alphaproteobacteria as graphically represented using Krona[5]. Decreased abundance of Archaea in the two deepest samples corresponds to an increase of Betaproteobacteria, Gammaproteobacteria, Bacteroidia, and Chlamydiae, and MetaPhlAn provided accurate assessments of certainty even for environments less well-covered by reference genomes. **B)** The five most abundant species according to a published blast-based approach[4] are compared to MetaPhlAn predictions. Both approaches consistently identified *Nitrosopumilus maritimus*, *Pelagibacter ubique*, and an as-yet-unsequenced clade in the *Pelagibacter* genus as the three most represented species in the marine oxygen minimum zone ecosystem.



Supplementary Fig. 7. Single-markers read counts for four representative species in the metagenomic stool samples from the Human Microbiome Project[6] show agreement among multiple within-species markers. For each of the four species, each box represents a sample (individual subject) and captures the distribution of read counts assigned to all single markers. These are normalized by the nucleotide length of the markers and ordered based on samples' median abundances for the species. Most samples show a tight distribution of each species' markers around a consensus relative abundance. Interestingly, nearly all outliers appear below the boxplots and represent underestimated marker abundances (i.e. partial false negatives). Cases where these cause extended interquartile ranges (e.g. *Prevotella copri* and *Ruminococcus bromii*) are consistently due to sample-specific strain variability, most likely due to differences in gene composition between the reference genomes and the genome in a subject's community. For example, several subjects' *P. copri* strains appear to match the sequence reference strain, but other subjects' strains lack large contiguous regions, resulting in unusually low counts for some markers. MetaPhlAn's classification strategy is robust to this, and additional sequenced genomes in the same species will automatically decrease the impact of this phenomenon. Moreover, the analysis of the distribution of specific markers across multiple samples can be exploited to characterize microbiomes sub-types at the strain level.

Supplementary Table 1: Genus level percentage precision for metagenomic reads (500,000 in total) sampled with an Illumina MAQ model (see methods) from 20 genomes belonging to species not included in the MetaPhlAn genomic repository. MetaPhlAn proved to outperform all alignment-free (PhyloPythiaS, Phymm), alignment-based (BlastN), and hybrid (PhymmBL) methods for 70% of newly classified microbes. Five of the six cases in which PhymmBL achieved better accuracies are due to MetaPhlAn's default BLASTN parameters, configured for "typical" microbes; increasing its sensitivity (word size 15, evalue 1e-15, reported as *MetaPhlAn strict*) allowed MetaPhlAN to outperform PhymmBL with only a slight computational overhead. For the remaining case (*Gluconobacter europaeus* NZ_CADR), MetaPhlAn detects a very low *Gluconobacter* abundance (<1%) but a high fraction (>90%) of an unclassified subclade in Acetobacteraceae, suggesting that this genome should be considered a genus distinct from *Gluconobacter*.

New Genome	New Species	Target Family	MetaPhlAn	MetaPhlAn strict	PhyloPythiaS	Phymm	BlastN	PhymmBL
NZ_BABW	Acetobacter aceti	Acetobacteraceae	**100.00**	86.76	0.00	1.35	2.41	2.75
NZ_AFBG	Acidovorax radicis	Comamonadaceae	11.98	**41.79**	2.04	5.96	13.47	13.66
NZ_CACP	Aeromonas caviae	Aeromonadaceae	**93.24**	83.25	3.98	19.60	81.02	81.26
NZ_AFBB	Dialister micraerophilus	Veillonellaceae	99.75	**99.86**	0.00	0.00	0.00	0.00
NZ_AEXB	Enterobacter mori	Enterobacteriaceae	49.36	**84.07**	0.06	10.35	76.22	76.70
NZ_ADLY	Enterococcus saccharolyticus	Enterococcaceae	**99.89**	91.39	0.00	1.54	8.73	8.88
NZ_CADR	Gluconacetobacter europaeus	Acetobacteraceae	0.83*	1.77*	0.00	16.86	67.10	**70.92**
NZ_AFBC	Haemophilus aegyptius	Pasteurellaceae	**99.94**	97.71	5.15	61.17	99.44	99.34
NZ_AFHS	Kingella kingae	Neisseriaceae	28.59	**83.37**	0.00	0.00	0.00	0.00
NZ_AEIZ	Leuconostoc fallax	Leuconostocaceae	48.58	**95.31**	0.50	9.04	24.67	28.51
NZ_AEOR	Leuconostoc lactis	Leuconostocaceae	58.50	**68.50**	0.72	9.31	38.81	40.83
NZ_AGAY	Neisseria shayeganii	Neisseriaceae	11.26	**27.39**	2.97	10.98	16.99	19.96
NZ_AFWQ	Neisseria weaveri	Neisseriaceae	**82.49**	73.89	7.23	24.09	26.68	35.63
NZ_AFHW	Paenibacillus elgii	Paenibacillaceae	0.00	**67.44**	1.89	5.55	18.78	20.81
NZ_AHBD	Pseudomonas psychrotolerans	Pseudomonadaceae	9.69	**72.35**	4.83	16.09	36.45	39.57
NZ_AHBW	Rhodococcus pyridinivorans	Nocardiaceae	**40.26**	12.26	7.14	8.71	24.47	29.16
NZ_AGIU	Saccharomonospora azurea	Pseudonocardiaceae	79.13	**83.90**	0.00	1.78	26.49	26.19
NZ_AEUV	Streptococcus criceti	Streptococcaceae	94.26	**98.15**	7.31	19.51	35.48	41.22
NZ_AEUZ	Streptococcus urinalis	Streptococcaceae	**88.87**	62.08	7.24	14.25	46.31	48.49
NZ_AFAJ	Vibrio rotiferianus	Vibrionaceae	**96.68**	88.09	2.90	25.44	81.68	82.92

*MetaPhlAn predicts this organism to be a genus among Acetobacteraceae distinct from *Gluconoacetobacter*. The low precision for *Gluconacetobacter* is thus likely to be due to a taxonomic misplacement of the new genome.

Supplementary Table 2 Family level precision for 20 genomes belonging to species without reference genomes (see Table 1 for details). In the great majority of cases (18 out of 20) MetaPhlAn outperform all existing methods using standard settings, whereas for the remaining 2 cases more sensitive parameters can be used to produce the best precision values.

New Genome	New Species	Target Family	MetaPhlAn	MetaPhlAn strict	PhyloPythiaS	Phymm	BlastN	PhymmBL
NZ_BABW	Acetobacter aceti	Acetobacteraceae	**100.00**	**100.00**	1.66	9.23	13.72	17.39
NZ_AFBG	Acidovorax radices	Comamonadaceae	61.12	**76.61**	5.09	21.70	48.28	51.82
NZ_CACP	Aeromonas caviae	Aeromonadaceae	**93.24**	83.36	4.42	19.64	81.05	81.30
NZ_AFBB	Dialister micraerophilus	Veillonellaceae	99.75	**99.90**	0.00	0.16	1.03	0.90
NZ_AEXB	Enterobacter mori	Enterobacteriaceae	50.02	**92.75**	4.54	45.21	89.78	90.84
NZ_ADLY	Enterococcus saccharolyticus	Enterococcaceae	**99.89**	91.39	0.00	2.20	13.66	13.56
NZ_CADR	Gluconacetobacter europaeus	Acetobacteraceae	93.33	**99.75**	1.97	20.40	69.18	73.40
NZ_AFBC	Haemophilus aegyptius	Pasteurellaceae	**99.97**	99.57	7.89	64.09	99.50	99.40
NZ_AFHS	Kingella kingae	Neisseriaceae	94.20	**100.00**	2.16	17.90	17.85	24.73
NZ_AEIZ	Leuconostoc fallax	Leuconostocaceae	59.17	**97.60**	0.68	9.60	25.87	29.87
NZ_AEOR	Leuconostoc lactis	Leuconostocaceae	58.50	**68.50**	0.82	9.77	39.70	41.69
NZ_AGAY	Neisseria shayeganii	Neisseriaceae	34.37	**99.01**	5.38	12.50	22.31	25.33
NZ_AFWQ	Neisseria weaveri	Neisseriaceae	96.10	**98.46**	7.86	24.23	28.10	36.62
NZ_AFHW	Paenibacillus elgii	Paenibacillaceae	0.00	**67.44**	2.48	5.78	19.72	21.65
NZ_AHBD	Pseudomonas psychrotolerans	Pseudomonadaceae	51.46	**77.36**	6.90	17.16	39.78	42.59
NZ_AHBW	Rhodococcus pyridinivorans	Nocardiaceae	**61.42**	18.17	8.63	10.20	28.65	33.59
NZ_AGIU	Saccharomonospora azurea	Pseudonocardiaceae	79.13	**83.90**	0.00	1.78	26.49	26.19
NZ_AEUV	Streptococcus criceti	Streptococcaceae	94.26	**98.15**	7.31	19.51	35.48	41.22
NZ_AEUZ	Streptococcus urinalis	Streptococcaceae	**88.87**	62.08	7.24	14.25	46.31	48.49
NZ_AFAJ	Vibrio rotiferianus	Vibrionaceae	**96.68**	88.09	2.90	25.44	81.68	82.92

1. M. Arumugam, J. Raes, E. Pelletier et al., *Nature* (2011).
2. G.D. Wu, J. Chen, C. Hoffmann et al., *Science* **334** (6052), 105 (2011).
3. J. Ravel, P. Gajer, Z. Abdo et al., *Proc Natl Acad Sci U S A* **108 Suppl 1**, 4680 (2011).
4. F.J. Stewart, O. Ulloa, and E.F. DeLong, *Environmental Microbiology* (2011).
5. B. Ondov, N. Bergman, and A. Phillippy, *BMC Bioinformatics* **12** (1), 385 (2011).
6. J. Peterson, S. Garges, M. Giovanni et al., *Genome Res* **19** (12), 2317 (2009).